

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

ในปัจจุบันการเก็บข้อมูลเอกสารภาษาไทยในรูปแบบอิเล็กทรอนิกส์มีปริมาณสูงขึ้นอย่างมากทำให้ความสนใจในการวิจัยเกี่ยวกับการออกแบบระบบการค้นคืนสารสนเทศมีความสำคัญอย่างมาก งานวิจัยทางด้านระบบการค้นคืนสารสนเทศเป็นงานวิจัยที่เกี่ยวข้องกับวิธีการจัดเก็บข้อมูล การจัดการภายในระบบและการค้นคืนเอกสารที่อยู่ในระบบ เอกสารที่จัดเก็บไว้จำนวนมากจะไร้ประโยชน์อย่างสิ้นเชิง ถ้าเอกสารเหล่านั้นไม่สามารถถูกค้นคืนได้ตรงกับความต้องการของผู้ใช้ ดังนั้นเป้าหมายสูงสุดของระบบการค้นคืนสารสนเทศจะคำนึงถึงความสามารถของระบบที่จัดหาสารสนเทศที่เกี่ยวข้องได้ตรงกับข้อคำถามของผู้ใช้

นักวิจัยหลายคนเสนอตัวแบบระบบการค้นคืนสารสนเทศไว้หลายแบบ (Frakes and Baeza-Yates, 1992; Baeza-Yates and Reibeiro-Neto, 1999; Berry and Browne, 2005) แต่โดยทั่วไปสามารถแบ่งตัวแบบได้ออกเป็นสองกลุ่ม ได้แก่ วิธีการเชิงคำหลัก (Keyword-oriented method) และวิธีการเชิงเมทริกซ์ (Matrix-oriented method) สำหรับวิธีการเชิงคำหลักจะใช้คำหลัก (Keyword) ซึ่งหมายถึง คำหรือกลุ่มคำเป็นข้อคำถาม (Query) ป้อนเข้าไปในระบบ แล้วคำหลักจะถูกนำไปจับคู่กับคำที่อยู่ในเอกสาร แล้วระบบจะคืนเอกสารที่มีคำหลักนั้นอยู่ให้แก่ผู้ใช้ระบบ ในขณะที่วิธีการเชิงเมทริกซ์จะมีการสร้างเมทริกซ์ขึ้นมาจากคำในเอกสารที่อยู่ในระบบ จากนั้นใช้วิธีการกระจายเมทริกซ์เป็นเมทริกซ์ย่อยเพื่อให้เห็นความสัมพันธ์ระหว่างคำและเอกสารออกมา (Ye, 2000)

การค้นคืนสารสนเทศด้วยวิธีการเชิงคำหลักมีปัญหาที่สำคัญคือ การค้นคืนอาจได้ผลไม่ถูกต้องตรงกับความต้องการของผู้ใช้ เนื่องจากความหมายของคำหลักหนึ่งอาจแสดงได้โดยคำหลายคำ (Synonymy) ส่งผลให้คำหลักในข้อคำถามไม่สามารถค้นคืนเอกสารที่เกี่ยวข้องออกมาได้ และคำหลักบางคำอาจมีหลายความหมาย (Polysemy) ทำให้คำหลักที่ใช้เป็นข้อคำถามไปดึงเอกสารที่ไม่เกี่ยวข้องออกมา การค้นคืนสารสนเทศด้วยวิธีการเชิงคำหลักจึงจะได้แต่เอกสารที่มีคำหลักตรงกับคำในข้อคำถามของผู้ใช้เท่านั้น ดังนั้นการค้นคืนสารสนเทศที่ดีจึงน่าจะเป็นการค้นคืนสารสนเทศจากความหมายที่มีอยู่ในเอกสาร

เทคนิคการค้นคืนสารสนเทศจากความหมายของเอกสารที่นิยมในปัจจุบัน คือ การจัดทำดัชนีหาความหมายแอบแฝง (Latent Semantic Indexing หรือ LSI) ซึ่งใช้วิธีการเชิงแมทริกซ์เพื่อจัดข้อบกพร่องของวิธีเชิงคำหลัก โดยการสร้างดัชนีจากความหมายของคำด้วยวิธีการทางสถิติมากกว่าที่จะใช้คำตามตัวอักษรในการค้นคืนสารสนเทศ การจัดทำดัชนีหาความหมายแอบแฝง (LSI) มีสมมุติฐานว่าเอกสารต่างๆจะมีความหมายซ่อนอยู่ในการใช้คำที่หลากหลาย เทคนิค LSI นี้ใช้การกระจายแมทริกซ์ (Matrix decomposition) ซึ่งวิธีการที่นิยมใช้เรียกว่า การกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition หรือ SVD) เพื่อดึงความสัมพันธ์ที่สำคัญระหว่างความหมายกับคำในเอกสาร (Berry , Dumais and O'Brien , 1995, p 573-595) แต่การกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (SVD) ต้องใช้เวลาในการกระจายแมทริกซ์มากและใช้พื้นที่ในการทำงานสูง ดังนั้นจึงไม่เหมาะสมในการนำไปใช้ระบบค้นคืนสารสนเทศที่มีเอกสารจำนวนมาก

ส่วนการกระจายแมทริกซ์แบบเซมิดีสครีต (Semidiscrete Matrix Decomposition หรือ SDD) เป็นเทคนิคที่พัฒนาต่อจากการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition หรือ SVD) โดยจะจำกัดค่าของแมทริกซ์เอกฐาน (Singular matrix) เป็นเลขจำนวนเต็ม ได้แก่ 0,1,-1 เท่านั้น ซึ่งทำให้ใช้พื้นที่ในการทำงานน้อยลง (Kolda and O'Leary, 1998, p. 322-346) เนื่องจากเทคนิค SDD เก็บข้อมูลส่วนหนึ่งในรูปแบบของเลขจำนวนเต็มแทนที่จะเก็บในรูปแบบเลขทศนิยม จึงทำให้ เทคนิค SDD มีความเหมาะสมกับการนำไปใช้กับฐานข้อมูลที่มีชุดเอกสารจำนวนมากมากกว่าเทคนิค SVD

งานวิจัยเกี่ยวกับการจัดทำดัชนีหาความหมายแอบแฝง (Latent Semantic Indexing หรือ LSI) ส่วนใหญ่เป็นงานวิจัยกับเอกสารภาษาอังกฤษ แต่ภาษาไทยซึ่งมีลักษณะเฉพาะที่ต่างจากภาษาอังกฤษหลายประการ เช่น โครงสร้างภาษา การรู้จักคำ การรู้จักประโยค และการทำความเข้าใจประโยค (Vongvipanond,1993, p. 519-545) มีงานวิจัยที่ทำการศึกษากำหนดดัชนีหาความหมายแอบแฝงอยู่บ้างโดยใช้เทคนิคการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition หรือ SVD) แต่เนื่องจากเทคนิค SVD จำเป็นต้องใช้พื้นที่การทำงานขนาดใหญ่และในการเปลี่ยนแปลงเอกสารในระบบการค้นคืนต้องใช้เวลาและมีขั้นตอนในการประมวลผลที่ยุ่งยากมาก ในขณะที่การกระจายแมทริกซ์แบบเซมิดีสครีต (Semidiscrete Matrix Decomposition หรือ SDD) ใช้เนื้อที่ในการประมวลผลน้อยกว่าและการปรับปรุงเปลี่ยนแปลงชุดเอกสารในระบบทำได้ง่ายกว่า วิทยานิพนธ์นี้จึงมุ่งหมายในการศึกษาเปรียบเทียบความสามารถในการค้นคืนสารสนเทศข้อความภาษาไทยโดยใช้เทคนิคการสร้าง

ดรรชนีหาความหมายแฝง ซึ่งมีการกระจายแมทริกซ์ระหว่างการกระจายแมทริกซ์แบบเซมิดีสครีต และการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว

1.2 วัตถุประสงค์ของงานวิจัย

(1) ศึกษาเปรียบเทียบความสามารถในการค้นคืนสารสนเทศข้อความภาษาไทย ระหว่างการกระจายแมทริกซ์แบบเซมิดีสครีต (Semidiscrete Matrix Decomposition หรือ SDD) และการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition หรือ SVD) ในชุดเอกสารข้อความภาษาไทยสาขาต่างๆ

(2) หาตัวแปรที่เหมาะสมที่ใช้ในการกระจายแมทริกซ์ ได้แก่ ตัวแปรแยกแมทริกซ์ (Splitting factor) และค่าขีดแบ่ง (Threshold value) สำหรับเอกสารภาษาไทยในหลายสาขาเพื่อใช้ในการปรับแต่งประสิทธิภาพในการค้นคืนสารสนเทศ

1.3 ขอบเขตของงานวิจัย

วิทยานิพนธ์นี้เป็นการศึกษาการค้นคืนสารสนเทศโดยเทคนิคการกระจายแมทริกซ์ในชุดเอกสารข้อความภาษาไทยซึ่งเป็นบทคัดย่อวิทยานิพนธ์จาก 4 สาขา ได้แก่ วิทยาการคอมพิวเตอร์ ประวัติศาสตร์ บรรณารักษศาสตร์ และรัฐศาสตร์

1.4 รายละเอียดของวิทยานิพนธ์

รายละเอียดของวิทยานิพนธ์ที่เหลืือสามารถแบ่งออกเป็นส่วนต่าง ๆ ซึ่งสามารถแยกอธิบายได้ดังหัวข้อต่อไปนี้

บทที่ 1 บทนำ ประกอบด้วย ความเป็นมาและความสำคัญของงานวิจัย วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย และรายละเอียดของวิทยานิพนธ์

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ประกอบด้วย ทฤษฎีต่างๆ ที่เกี่ยวข้องกับงานวิจัยนี้ และผลงานวิจัยที่นำเทคนิคการกระจายแมทริกซ์ไปประยุกต์ใช้กับการค้นคืนสารสนเทศ และประยุกต์ใช้กับภาษาไทย

บทที่ 3 วิธีการดำเนินงานวิจัย ประกอบด้วยรายละเอียดของข้อมูลที่ใช้ในการวิจัย เครื่องมือที่ใช้ ขั้นตอนของการวิจัย และวิธีวัดผลการวิจัย

บทที่ 4 ผลการทดลอง แสดงผลการทดลองปัจจัยที่ใช้ในการทดสอบ

บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ ประกอบด้วย สรุปและอภิปรายผลการ ศึกษาวิจัย และข้อเสนอแนะต่าง ๆ

